Attorney Docket No. 122.1580

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:

Xiaohong Huang, et al.

Application No.: 10/768,178 Group Art Unit:

Filed: February 2, 2004 Examiner:

For: AN APPARATUS AND METHOD FOR EXTRACTING INFORMATION FROM A FORMATTED DOCUMENT

SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN APPLICATION IN ACCORDANCE WITH THE REQUIREMENTS OF 37 C.F.R. § 1.55

Commissioner for Patents
PO Box 1450
Alexandria, VA 22313-1450

Sir:

In accordance with the provisions of 37 C.F.R. § 1.55, the applicant(s) submit(s) herewith a certified copy of the following foreign application:

Japanese and Chinese Patent Application No(s). PCT/JP02/07983 and 01 1 23845.3

Filed: August 5, 2002 and August 3, 2001

It is respectfully requested that the applicant(s) be given the benefit of the foreign filing date(s) as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. § 119.

Respectfully submitted,

STAAS & HALSEY LLP

Date: March 11, 2004 By: _____

H. J. Staas
Registration No. 22,010

1201 New York Ave, N.W., Suite 700
Washington, D.C. 20005
Telephone: (202) 434-1500
Facsimile: (202) 434-1501

日 本 国 特 許 庁

JAPAN PATENT OFFICE

別紙添付の書類は下記の出願書類の謄本に相違ないことを証明する。
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 Date of Application:	２００２年　８月　５日
出願番号 Application Number:	ＰＣＴ／ＪＰ０２／０７９８３
出願人 Applicant (s):	FUJITSU LIMITED HUANG, Xiaohong XU, Guowei

２００４　年　３　月　４　日

特許庁長官
Commissioner,
Japan Patent Office　今井康夫

出証平　１６－５００６８

HOME COPY

PCT REQUEST

K869-PCT

Original (for SUBMISSION) - printed on 05.08.2002 05:04:11 PM

0	For receiving Office use only	
0-1	International Application No.	PCT/JP 02/07983
0-2	International Filing Date	05.08.02
0-3	Name of receiving Office and "PCT International Application"	PCT International Application JAPAN PATENT OFFICE

0-4	Form - PCT/RO/101 PCT Request	
0-4-1	Prepared using	PCT-EASY Version 2.92 (updated 01.01.2002)
0-5	Petition The undersigned requests that the present international application be processed according to the Patent Cooperation Treaty	
0-6	Receiving Office (specified by the applicant)	Japan Patent Office (RO/JP)
0-7	Applicant's or agent's file reference	K869-PCT
I	Title of invention	AN APPARATUS AND METHOD FOR EXTRACTING INFORMATION FROM A FORMATTED DOCUMENT
II	Applicant	
II-1	This person is:	applicant only
II-2	Applicant for	all designated States except US
II-4	Name	FUJITSU LIMITED
II-5	Address:	1-1, Kamikodanaka 4-chome, Nakahara-ku, Kawasaki-shi, Kanagawa 211-8588 Japan
II-6	State of nationality	JP
II-7	State of residence	JP
III-1	Applicant and/or inventor	
III-1-1	This person is:	applicant and inventor
III-1-2	Applicant for	US only
III-1-4	Name (LAST, First)	HUANG, Xiaohong
III-1-5	Address:	C/O Fujitsu Research & Development Center Co., Ltd. Room B1003, Eagle Run Plaza No.26 Xiaoyun Road, Chaoyan District Beijing 100016, P.R. China China
III-1-6	State of nationality	CN
III-1-7	State of residence	CN

RO

RO

PCT REQUEST K869-PCT

Original (for **SUBMISSION**) - printed on 05.08.2002 05:04:11 PM

III-2	**Applicant and/or inventor**	
III-2-1	This person is:	applicant and inventor
III-2-2	Applicant for	US only
III-2-4	Name (LAST, First)	XU, Guowei
III-2-5	Address:	C/O Fujitsu Research & Development Center Co., Ltd. Room B1003, Eagle Run Plaza No.26 Xiaoyun Road, Chaoyan District Beijing 100016, ~~P.R. China~~ China
III-2-6	State of nationality	CN
III-2-7	State of residence	CN
IV-1	**Agent or common representative; or address for correspondence** The person identified below is hereby/has been appointed to act on behalf of the applicant(s) before the competent International Authorities as:	agent
IV-1-1	Name (LAST, First)	ISHIDA, Takashi
IV-1-2	Address:	A. AOKI, ISHIDA & ASSOCIATES Toranomon 37 Mori Bldg., 5-1, Toranomon 3-chome, Minato-ku, Tokyo 105-8423 Japan
IV-1-3	Telephone No.	03-5470-1900
IV-1-4	Facsimile No.	03-5470-1911
IV-2	**Additional agent(s)**	additional agent(s) with same address as first named agent
IV-2-1	Name(s)	TSURUTA, Junichi; TSUCHIYA, Shigeru; NISHIYAMA, Masaya; HIGUCHI, Sotoji
V	**Designation of States**	
V-1	Regional Patent (other kinds of protection or treatment, if any, are specified between parentheses after the designation(s) concerned)	EP: AT BE CH&LI CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE TR and any other State which is a Contracting State of the European Patent Convention and of the PCT
V-2	National Patent (other kinds of protection or treatment, if any, are specified between parentheses after the designation(s) concerned)	JP US

RO

PCT REQUEST K869-PCT

Original (for **SUBMISSION**) - printed on 05.08.2002 05:04:11 PM

V-5	**Precautionary Designation Statement** In addition to the designations made under items V-1, V-2 and V-3, the applicant also makes under Rule 4.9(b) all designations which would be permitted under the PCT except any designation(s) of the State(s) indicated under item V-6 below. The applicant declares that those additional designations are subject to confirmation and that any designation which is not confirmed before the expiration of 15 months from the priority date is to be regarded as withdrawn by the applicant at the expiration of that time limit.		
V-6	**Exclusion(s) from precautionary designations**	NONE	
VI-1	**Priority claim of earlier national application**		
VI-1-1	Filing date	03 August 2001 (03.08.2001)	
VI-1-2	Number	01123845.3	
VI-1-3	Country	CN	
VII-1	**International Searching Authority Chosen**	European Patent Office (EPO) (ISA/EP)	
VIII	**Declarations**	Number of declarations	
VIII-1	Declaration as to the identity of the inventor	-	
VIII-2	Declaration as to the applicant's entitlement, as at the international filing date, to apply for and be granted a patent	-	
VIII-3	Declaration as to the applicant's entitlement, as at the international filing date, to claim the priority of the earlier application	-	
VIII-4	Declaration of inventorship (only for the purposes of the designation of the United States of America)	-	
VIII-5	Declaration as to non-prejudicial disclosures or exceptions to lack of novelty	-	
IX	**Check list**	number of sheets	electronic file(s) attached
IX-1	Request (including declaration sheets)	4	-
IX-2	Description	6	-
IX-3	Claims	3	-
IX-4	Abstract	1	-
IX-5	Drawings	5	-
IX-7	TOTAL	19	
	Accompanying items	paper document(s) attached	electronic file(s) attached
IX-8	Fee calculation sheet	✓	-
IX-17	PCT-EASY diskette	-	Diskette
IX-18	Other (specified):	patent revenue stamps	-
IX-19	**Figure of the drawings which should accompany the abstract**	1	

K869-PCT

PCT REQUEST

Original (for **SUBMISSION**) - printed on 05.08.2002 05:04:11 PM

IX-20	Language of filing of the international application	English
X-1	Signature of applicant, agent or common representative	Takashi Ishida
X-1-1	Name (LAST, First)	ISHIDA, Takashi
X-2	Signature of applicant, agent or common representative	Junichi Tsuruta
X-2-1	Name (LAST, First)	TSURUTA, Junichi
X-3	Signature of applicant, agent or common representative	Tsuchiya Shigeru
X-3-1	Name (LAST, First)	TSUCHIYA, Shigeru
X-4	Signature of applicant, agent or common representative	M. Nishiyama
X-4-1	Name (LAST, First)	NISHIYAMA, Masaya
X-5	Signature of applicant, agent or common representative	Sotoji Higuchi
X-5-1	Name (LAST, First)	HIGUCHI, Sotoji

FOR RECEIVING OFFICE USE ONLY

10-1	Date of actual receipt of the purported international application	05.08.02
10-2	Drawings:	
10-2-1	Received	
10-2-2	Not received	
10-3	Corrected date of actual receipt due to later but timely received papers or drawings completing the purported international application	
10-4	Date of timely receipt of the required corrections under PCT Article 11(2)	
10-5	International Searching Authority	ISA/EP
10-6	Transmittal of search copy delayed until search fee is paid	✓

FOR INTERNATIONAL BUREAU USE ONLY

11-1	Date of receipt of the record copy by the International Bureau	

Description

An apparatus and method for extracting information from a formatted document

Technical Field

The present invention in general relates to an apparatus and method for extracting information from an input formatted document, and in particular , to an apparatus and method for automatically extracting special character strings from an input formatted document, for example from web pages of online sale.

Background Art

It is known in the art an apparatus for extracting text information from a document, such as the technology disclosed in S. Soderland's article entitled of "Learning to Extract Text-base Information from the World Wide Web" (Proc. 3rd Intl Conf. On Knowledge Discovery and Data Mining (KDD-97)). In such an apparatus, the special character strings are distinguished by means of the character strings being the function of attribute names (e.g. "goods names") and placed before the special character strings, and are then extracted.

In the prior art apparatus, since the special character strings are distinguished and extracted by means of the character strings being the function of attribute names (such as "goods names", etc.) and placed before the special character strings, it is effective when the attribute names such as "goods names" as well as the attribute values such as "monogram accessory pouch" are available. However, the documents such as the web pages of Internet have various formats. Therefore, there is a situation that the attribute names fail to be provided. For example, only the character strings "monogram accessory pouch" are provided. In the case that the attribute names are not provided, the special character strings can not be extracted by means of the above-mentioned technology. Moreover, in the present technology, the machine can not extract the special character strings automatically, if samples are not provided manually for the machine.

Summary of the Invention

To solve the above problems, the present invention is attained. Therefore, an object of the invention is to provide an apparatus and a method for automatically special character strings from an input formatted document.

In order to accomplish the object of the invention, there is provided an apparatus for extracting text information from an input formatted document, comprising: an input unit for inputting a formatted document; a unit for analyzing the input formatted document and saving the particular typographic information; a unit for identifying special character strings by means of the typographic information such as font size, character font, color, etc.; a unit for extracting the identified special character strings; and an output unit for outputting the extracted character strings.

According to another aspect of the invention, a method for extracting information from a formatted document is provided, which comprises the following steps: inputting a formatted document; analyzing the input formatted document and saving the particular typographic information; identifying special character strings by means of the typographic information such as font size, character font, color, etc.; extracting the identified special character strings; and outputting the extracted character strings.

According to the invention, the operations of analyzing the input formatted document, identifying special character strings by means of the typographic information such as font size, character font, color, etc and extracting the special character strings enable to automatically extract special character strings from the input formatted document and considerably increase the accuracy of extraction. Moreover, the prior apparatus requires to manually input samples for memory, while the apparatus according to the invention can automatically carry out the determination and extraction with respect to different types of the formatted document without inputting the samples.

Brief Description of the Drawings

FIG. 1 is a structural block chart of the apparatus for extracting information from a formatted document according to the invention.

FIG. 2 is document data and a flowchart illustrating a first embodiment of the invention.

FIG. 3 document data and a flowchart illustrating a second embodiment of the invention.

FIG. 4 is document data and a flowchart illustrating a third embodiment of the invention.

FIG. 5 is document data and a flowchart illustrating a fourth embodiment of the invention.

Best Mode for Carrying out the Invention

As shown in figure 1, there is a structural block chart of the apparatus for extracting information from a formatted document according to the invention.

In the extraction apparatus for extracting information from a formatted document as shown in figure 1, numeral 1 indicates an input unit for inputting a formatted document; 2 indicates a unit for analyzing the input formatted document through a certain method and saving the particular typographic information, 3 is a unit for identifying special character strings on the basis of the analysis result by means of the typographic information such as font size, character font, color, etc., 5 is a unit for extracting the identified special character strings, and 5 is an output unit for outputting the extracted character strings.

Next, the actions of the apparatus according to the invention will be described in detail with reference to figures 2 to 5 by an example of extracting special character strings from HTML document.

Example 1

FIG. 2 is document data and a flowchart illustrating a first embodiment of the invention, wherein figure 2 (a) is sale information which are obtained from a certain network and are a document in the form of HTML, figure 2(b) is HTML source file of the information shown in figure 2(a), figure 2(c) is a flowchart illustrating the actions of extracting information in example 1.

Next, the flow of information extraction steps in example 1 is described as follows. In step 101, HTML source file as shown in figure 2 (b) is inputted. In step 102, the thus input HTML source file is analyzed so as to find typographic information. Then, in steps 103-107, the special character strings are extracted.

At first, in step 103, the character strings to be discriminated are determined on the basis of the result obtained in step 102. Then, in step 104, a decision should be made on whether the font size of the character strings determined in step 103 is the biggest one with respect to the surrounding character strings. If it is not, then turns to the step 106. In step 106, a decision is made on whether the typographic information of said character strings is beyond the range of the preset values. If it is yes, then goes into step 107 in which the information extraction action is ended. If it is not,

then returns to step 103 and thus determine the next character strings to be discriminated.

If the decision in step 104 is "yes", that is, the typographic information of the character string "Windows Operation and Application Technology(second version)" in example 1 is (FONT size=5) and is the biggest among the surrounding character strings, it is determined as special typographic information. Then, goes into step 105, in which the character string "Windows Operation and Application Technology(second version)" is determined as special character strings, i.e., goods name.

Using the information extraction apparatus according to the present embodiment, the special character string is enable to be automatically extracted from the input formatted document by discriminating it via typographic information such as font size.

Example 2

FIG. 3 is document data and a flowchart illustrating the second embodiment of the invention, wherein figure 3(a) is sale information which are obtained from a certain network and are a document in the form of HTML, figure 3(b) is HTML source file of the information shown in figure 3(a), figure 3(c) is a flowchart illustrating the actions of extracting information in example 2.

Next, the information extraction process in example 2 is described as follows. For clarity of illustration, the same steps as those described in the above example 1 are omitted, and only the different steps are described as below.

In step 204, a decision should be made on whether, for example, the font of the character string determined in step 203 is different from the surrounding character strings. If the decision in step 204 is "yes", that is, the typographic information of the character string "Windows Operation and Application Technology(second version)" in example 2 is (FONT "Chinese regular script" and the color is red(color = # ff0000)) and is particularly different from the surrounding character strings, it is determined as special typographic information. Then, goes into step 205, in which the character string "Windows Operation and Application Technology(second version)" is discriminated as special character strings, i.e., goods name.

Using the information extraction apparatus according to the

present embodiment, the special character string is enable to be automatically extracted from the input formatted document by discriminating it via typographic information such as font and color.

Example 3

FIG. 4 is document data and a flowchart illustrating the third embodiment of the invention, wherein figure 4(a) is sale information which are obtained from a certain network and are a document in the form of HTML, figure 4(b) is HTML source file of the information shown in figure 4(a), figure 4(c) is a flowchart illustrating the actions of extracting information in example 3.

Next, information extraction process in example 3 is described in detail. For clarity of illustration, the same steps as those described in the above example 1 are omitted, and only the different steps are described as below.

In step 304, a decision should be made on whether, for example, the font of the character string determined in step 303 is different from the surrounding character strings. If the decision in step 304 is "yes", that is, the typographic information of the character string "Windows Operation and Application Technology(second version)" in this example is (FONT "Chinese regular script" and boldface (<B><FONT...</B>)) and is particularly different from the surrounding character strings, it is determined as special typographic information. Then, goes into step 305, in which the character string "Windows Operation and Application Technology(second version)" is discriminated as special character strings, i.e., goods name.

Using the information extraction apparatus according to the present embodiment, the special character string is enable to be automatically extracted from the input formatted document by discriminating it via typographic information such as font and boldface.

Example 4

FIG. 5 is document data and a flowchart illustrating the fourth embodiment of the invention, wherein figure 5(a) is sale information which are obtained from a certain network and are a document in the form of HTML; figure 5(b) is HTML source file of the information shown in figure 5(a); figure 5(c) is a flowchart illustrating the actions of extracting information in example 4.

Next, information extraction process in example 4 is described in detail. For clarity of illustration, the same steps as those described in the above example 1 are omitted, and only the different steps are described as below.

In step 404, a decision should be made on whether, for example, the font of the character string determined in step 403 is different from the surrounding character strings. If the decision in step 404 is "yes", that is, the typographic information of the character string "Windows Operation and Application Technology(second version)" in this example is (red color (color = #ff0000) and boldface) and is particularly different from the surrounding character strings, it is determined as special typographic information. Then, goes into step 405, in which the character string "Windows Operation and Application Technology(second version)" is discriminated as special character strings, i.e., goods name.

Using the information extraction apparatus according to the this embodiment, the special character string is enable to be automatically extracted from the input formatted document by discriminating it via typographic information such as color and boldface.

It should be understood, however, that the above disclosure with respect to the examples 1-4 is illustrative only, other than any limitation to the present invention. Any modifications and variations to the embodiments 1-4 of the invention may be made without departing from the spirit and the protection scope of the invention defined by the appended claims. For example, proper combination and variation of the embodiments 1-4 can be made and can obtain the same effect of the invention, i.e., automatically extracting special character strings.

Claims

1. An apparatus for extracting information from a formatted document, comprising: an input unit (1) for inputting a formatted document; a unit (2) for analyzing the input formatted document and saving the particular typographic information; a unit (3) for identifying special character strings on the basis of the analysis result by means of the typographic information such as font size, character font, color, etc., a unit (4) for extracting the identified special character strings; and an output unit (5) for outputting the extracted character strings.

2. The apparatus for extracting information from a formatted document according to claim 1, wherein said unit (3) for identifying special character strings determines a certain character string as a special one on the basis of the typographic information of said formatted document when the typographic information of said character string is determined as a special typographic information.

3. The apparatus for extracting information from a formatted document according to claim 1 or 2, wherein said formatted document is HTML document, and said unit (3) for identifying special character strings a certain character string as a special one on the basis of the analyzing results with respect to said HTML document when the font size of said character string is determined to be the biggest one among the surrounding character strings.

4. The apparatus for extracting information from a formatted document according to claim 1 or 2, wherein said formatted document is HTML document, and said unit (3) for identifying special character strings determines a certain character string as a special one on the basis of the analyzing results with respect to said HTML document when the color and the font of said character string is determined to be a special one among the surrounding character strings.

5. The apparatus for extracting information from a formatted document according to claim 1 or 2, wherein said formatted document is HTML document, and said unit (3) for identifying special character strings determines a certain character string as a special one on the basis of the analyzing results with respect to said HTML document when the font of said character string is determined to be different from the surrounding character strings and said character string to be boldface.

6. The apparatus for extracting information from a formatted document according to claim 1 or 2, wherein said formatted document is HTML document, and said unit (3) for identifying special character strings determines a certain character string as a special one on the basis of the analyzing results with respect to said HTML document when the color of said character string is determined to be different from the surrounding character strings and said character string to be boldface.

7. A method for extracting information from a formatted document, comprising the following steps ; inputting a formatted document, analyzing the input formatted document and saving the particular typographic information; identifying special character strings on the basis of the analysis result by means of the typographic information such as font size, character font, color, etc.; extracting the identified special character strings; and outputting the extracted character strings.

8. The method according to claim 8, wherein in the step of identifying special character string, a celtain character string is determined as a special one on the basis of the typographic information of said formatted document when the typographic information of said character string is determined as a special typographic information.

9. The method according to claim 7 or 8, wherein said formatted document is HTML document, and in the step of identifying special character string, a certain character string is determined as a special one on the basis of the analyzing results with respect to said HTML document when the font size of said character string is determined to be the biggest one among the surrounding character strings.

10. The method according to claim 7 or 8, wherein said formatted document is HTML document, and in the step of identifying special character string, a certain character string is determined as a special one on the basis of the analyzing results with respect to said HTML document when the color and the font of said character string is determined to be a special one among the surrounding character strings.

11. The method according to claim 7 or 8, wherein said formatted document is HTML document, and in the step of identifying special

character string, a certain character string is determined as a special one on the basis of the analyzing results with respect to said HTML document when the font of said character string is determined to be different from the surrounding character strings and said character string to be boldface.

12. The method according to claim 7or 8, wherein said formatted document is HTML document, and in the step of identifying special character string, a certain character string is determined as a special one on the basis of the analyzing results with respect to said HTML document when the color of said character string is determined to be different from the surrounding character strings and said character string to be boldface.

Abstract

The present invention discloses an apparatus for extracting information from a formatted document, comprising: an input unit (1) for inputting a formatted document; a unit (2) for analyzing the input formatted document and saving the particular typographic information, a unit (3) for identifying special character strings on the basis of the analysis result by means of the typographic information such as font size, character font, color, etc.; a unit (4) for extracting the identified special character strings; and an output unit (5) for outputting the extracted character strings. When the typographic information of a certain character string is determined as a special typographic information, said character string is determined to be special character string. Thus, the present apparatus is able to automatically extract information from different types of format documents.

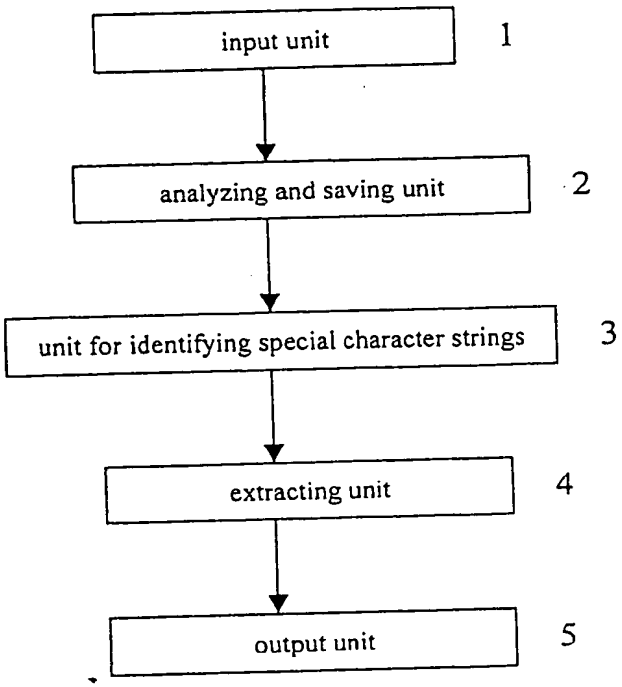

input unit
1
analyzing and saving unit
2
unit for identifying special character strings
3
extracting unit
4
output unit
5

FIG. 1

FIG.2

(a)

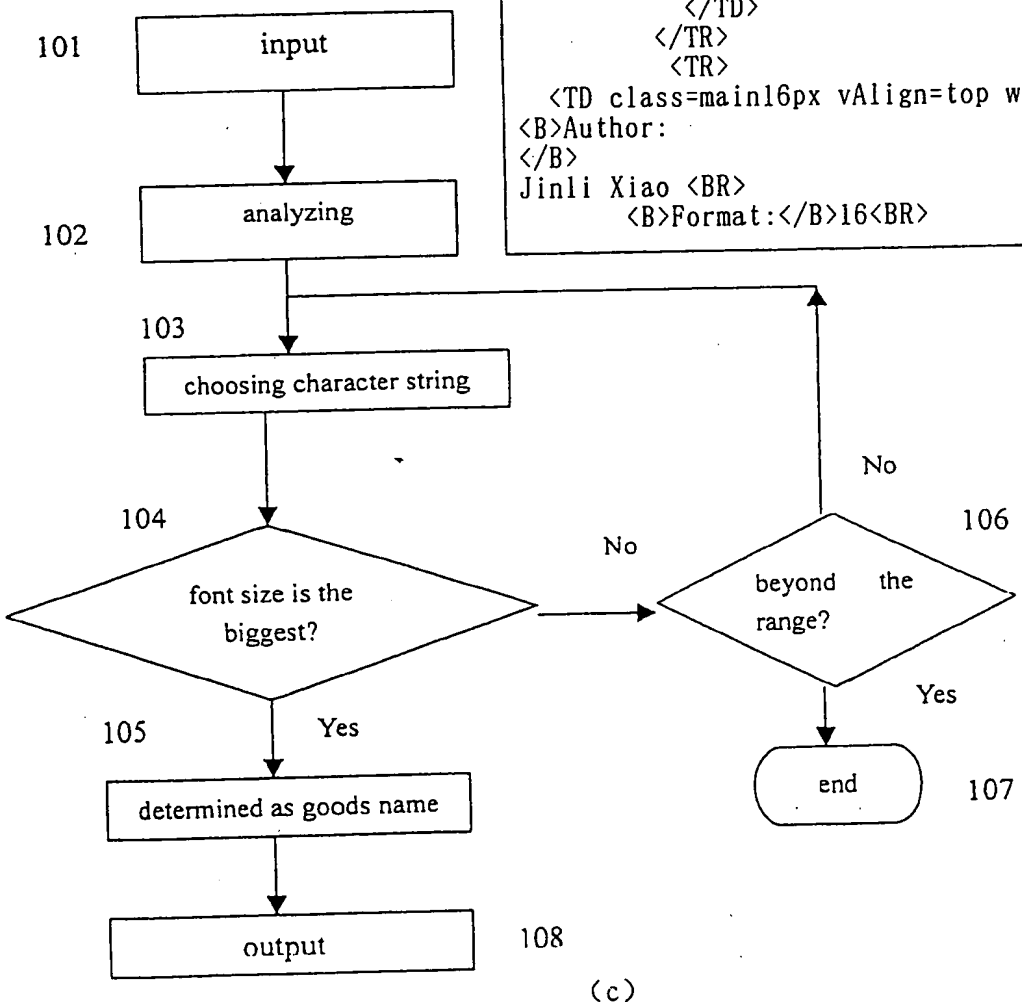
Information of Books

Windows Operation and Application Technology (second version)

Author:Jinli Xiao
Format:16
Setting:
Pages:240
Publishing Company:Electronical Industrial Publication Company
Publication date:August 1, 1997
ISBN:7505340492

(b)

```
<FONT size=5>Windows Operation and
Application Technology (second version)
        
     </FONT></P>
           <P> </P>
         </TD>
       </TR>
        <TR>
  <TD class=main16px vAlign=top width="18%">
<B>Author:
</B>
Jinli Xiao <BR>
       <B>Format:</B>16<BR>
```



(c)

FIG. 3

(a)

Information of Books

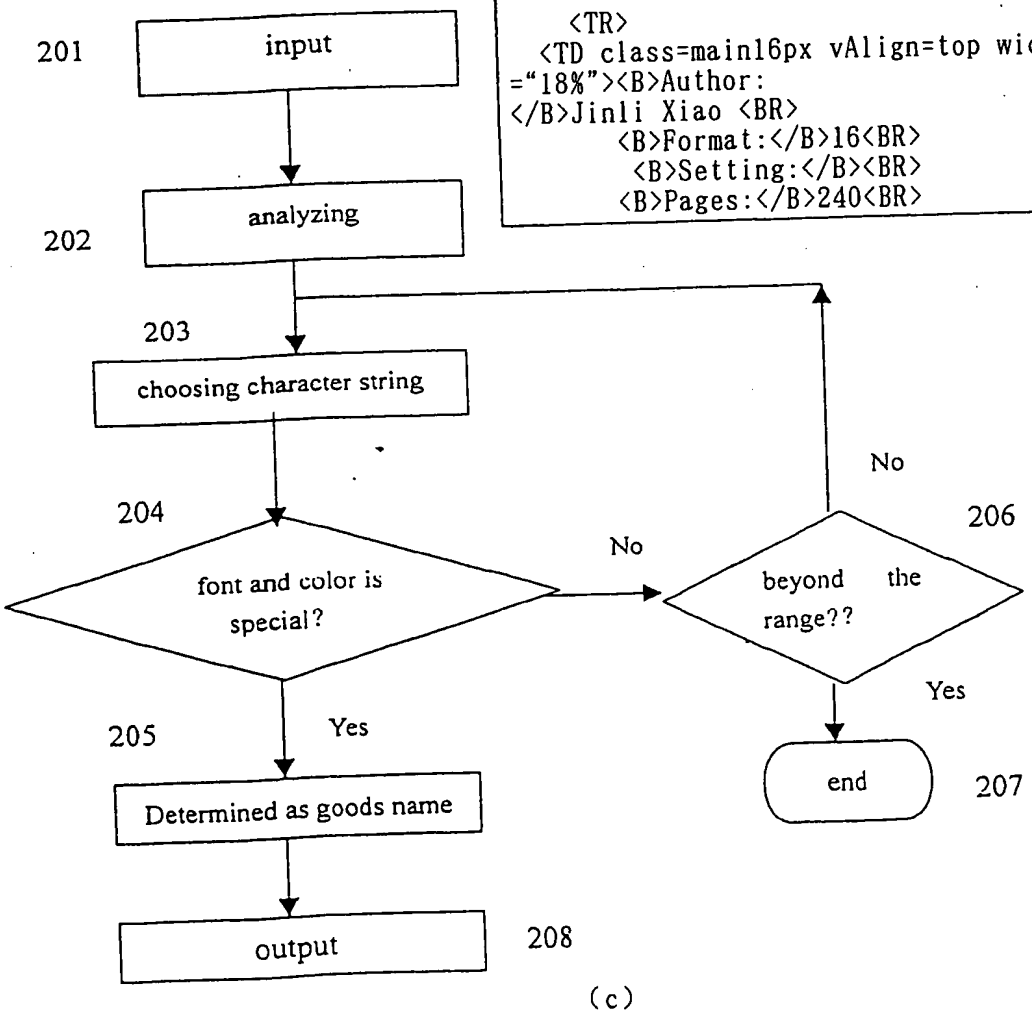
Windows Operation and Application Technology (second version)

Author:Jinli Xiao
Format:16
Setting:
Pages:240
Publishing Company:Electronical Industrial Publication Company
Publication date:August 1, 1997
ISBN:7505340492

(b)

```
P><font face="Chinese regular script"
color=#ff0000>Windows Operation and
Application Technology (second version)
      </font></P>
            <P> </P>
          </TD>      </TR>

   <TR>
  <TD class=main16px vAlign=top width
="18%"><B>Author:
</B>Jinli Xiao <BR>
       <B>Format:</B>16<BR>
        <B>Setting:</B><BR>
       <B>Pages:</B>240<BR>
```



(c)

FIG. 4

(a)

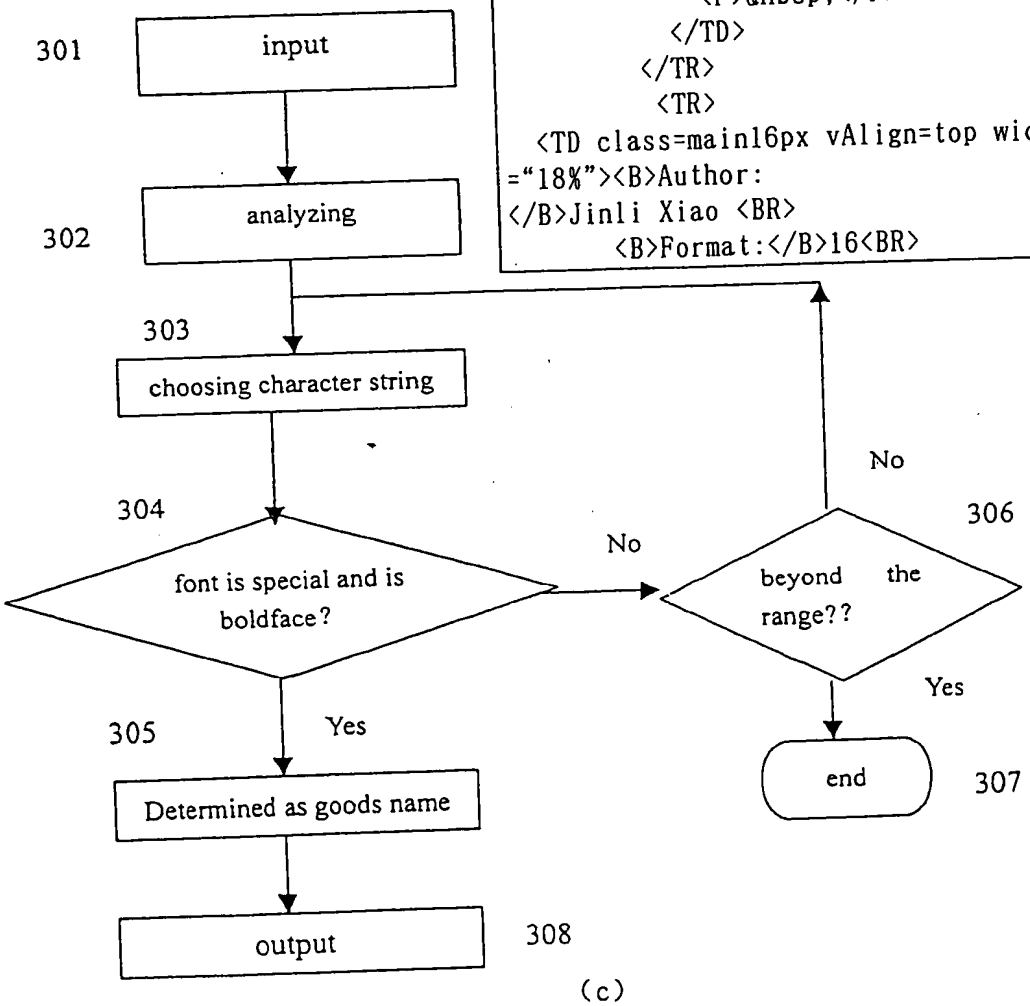
Information of Books

Windows Operation and Application Technology (second version)

Author:Jinli Xiao
Format:16
Setting:
Pages:240
Publishing Company:Electronical Industrial Publication Company
Publication date:August 1, 1997
ISBN:7505340492

(b)

```
<P><B><FONT face="Chinese regular script">
Windows Operation and Application
Technology (second version)
          
      </FONT></B></P>
            <P> </P>
          </TD>
        </TR>
         <TR>
  <TD class=main16px vAlign=top width
="18%"><B>Author:
</B>Jinli Xiao <BR>
       <B>Format:</B>16<BR>
```



(c)

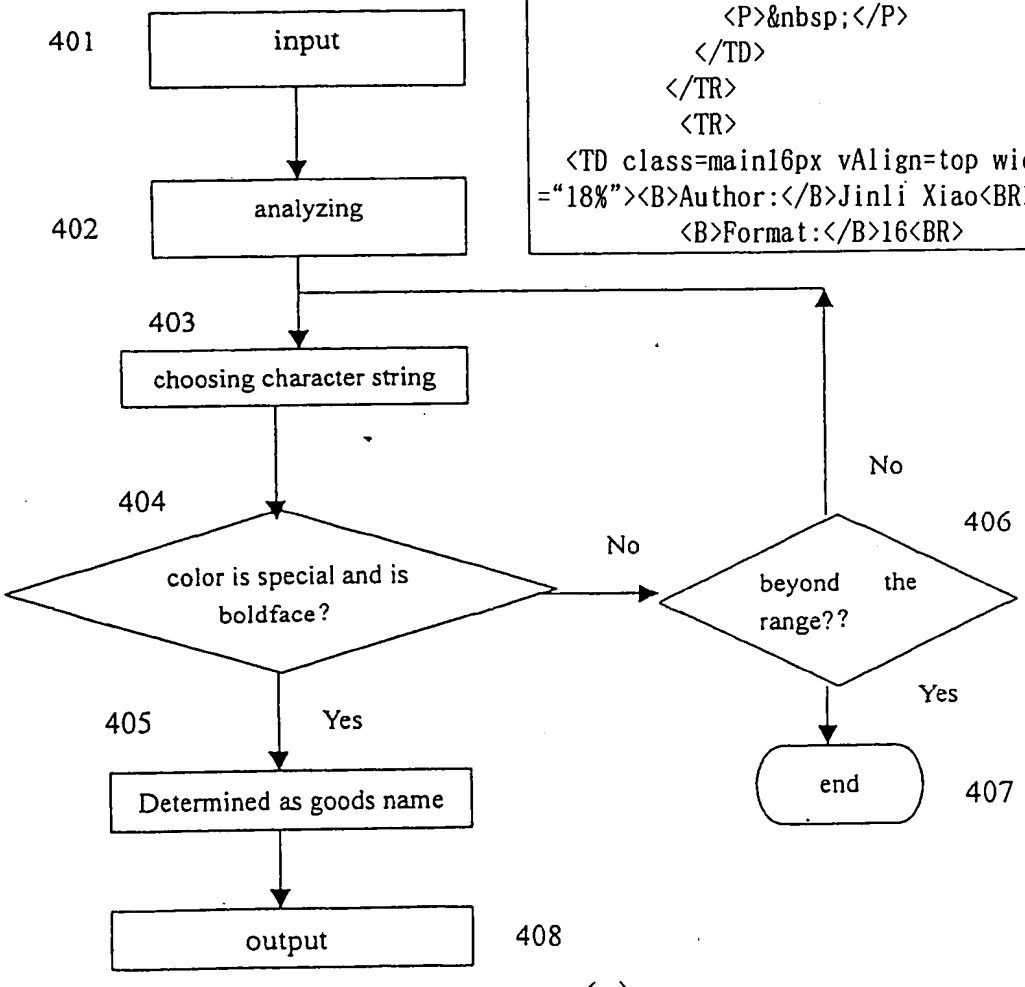
FIG. 5

(a)

Information of Books
Windows Operation and Application Technology (second version)
Author:Jinli Xiao
Format:16
Setting:
Pages:240
Publishing Company:Electronical Industrial Publication Company
Publication date:August 1, 1997
ISBN:7505340492

(b)

```
<P><B><FONT color=#ff0000>Windows
Operation and Application Technology
(second version)
  
</FONT></B></P>
<P> </P>
</TD>
</TR>
<TR>
<TD class=main16px vAlign=top width
="18%"><B>Author:</B>Jinli Xiao<BR>
<B>Format:</B>16<BR>
```



(c)